A New Method for Recognizing Cytokines Based on Feature Combination and a Support Vector Machine Classifier

Conclusion –

We propose a method which can recognize cytokines from a dataset that is highly skewed on the ratio of 1:9, i.e., cytokines are very less in comparison of non-cytokines. We extract the features based on four different methods and then perform classifier selection on obtained features. The first way to obtain features based on amino acid composition (AAC) of the 20-dimensional features vector. In a second way, we try each possible combination of a polypeptide based on the physicochemical properties of proteins to make a feature vector of 400 lengths. The third techniques are used to extract the features is based on position-specific score matrices (PSSM) which can be computed by PSI-BLAST software [1]. And finally, we compute features based on structure probability matrices (SPM) and secondary structure sequences (SSS). We obtained a total of 868 features, and we select 448 features. We use Support Vector Machine (SVM) with linear kernel function the reason of choosing Support Vector Machine (SVM) as a classifier because it is more applicable on binary classification, and it is consistent with our problem, and we get 93.25% accuracy with that. We perform 10 cross-validations to determine the generalized ability of our method.
Limitation of our method is that our model can only apply for two class problem. But we can overcome this problem by extending our method using principal component analysis (PCA) method for features selection and then apply multi-class support vector machine (MSVM) as a classifier to extend it for multiclass classification.
We chose Fpsp with the extracted features of length 448 with the linear kernel function of Support Vector Machine as our classifier. Our method is well performed as comparisons to others method that has been proposed yet because our result get at least a 2.07% and at most a 5.96% increase in accuracy by checking the 10 cross-validations, at least a 0.42% and at most a 14.51% increase in sensitivity.

References –

1. Altschul, S.F.; Madden, T.L.; Schäffer, A.A.; Zhang, J.; Zhang, Z.; Miller,W.; Lipman, D.J. Gapped blast and psi-blast: A new generation of protein database search programs. Nucleic Acids Res. 1997, 25, 3389–4002. [CrossRef] [PubMed]